3D convolutional neural network for differentiating pre-invasive lesions from invasive adenocarcinomas appearing as ground-glass nodules with diameters ≤3 cm using HRCT

Shengping Wang¹,², Rui Wang¹, Shengjian Zhang¹,², Ruimin Li¹,², Yi Fu¹,², Xiangjie Sun²,⁴, Yuan Li²,⁴, Xing Sun³, Xinyang Jiang³, Xiaowei Guo³, Xuan Zhou³, Jia Chang³, Weijun Peng¹,²

¹Department of Radiology, Fudan University Shanghai Cancer Center, Shanghai 200032, China; ²Department of Oncology, Shanghai Medical College, Fudan University, Shanghai 200032, China; ³Tencent Youtu Lab, Shanghai 200050, China; ⁴Department of Pathology, Fudan University Shanghai Cancer Center, Shanghai 200032, China

Correspondence to: Weijun Peng. Department of Radiology, Fudan University Shanghai Cancer Center, No. 270 Dong An Road, Xuhui District, Shanghai 200032, China. Email: cjr.pengweijun@vip.163.com.

Background: Identification of pre-invasive lesions (PILs) and invasive adenocarcinomas (IACs) can facilitate treatment selection. This study aimed to develop an automatic classification framework based on a 3D convolutional neural network (CNN) to distinguish different types of lung cancer using computed tomography (CT) data.

Methods: The CT data of 1,545 patients suffering from pre-invasive or invasive lung cancer were collected from Fudan University Shanghai Cancer Center. All of the data were preprocessed through lung mask extraction and 3D reconstruction to adapt to different imaging scanners or protocols. The general flow for the classification framework consisted of nodule detection and cancer classification. The performance of our classification algorithm was evaluated using a receiver operating characteristic (ROC) analysis, with diagnostic results from three experienced radiologists.

Results: The sensitivity, specificity, accuracy, and AUC (area under the ROC curve) values of our proposed automatic classification method were 88.5%, 80.1%, 84.0%, and 89.2%, respectively. The results of the CNN classification method were compared to those of three experienced radiologists. The AUC value of our method (AUC =0.892) was higher than those of all radiologists (radiologist 1: 80.5%; radiologist 2: 83.9%; and radiologist 3: 86.7%).

Conclusions: The 3D CNN-based classification algorithm is a promising tool for the diagnosis of pre-invasive and invasive lung cancer and for the treatment choice decision.

Keywords: Pre-invasive lesions (PILs); invasive adenocarcinomas (IACs); 3D convolution neural network (3D CNN); automatic diagnosis

Submitted Mar 30, 2018. Accepted for publication Jun 06, 2018.
doi: 10.21037/qims.2018.06.03
View this article at: http://dx.doi.org/10.21037/qims.2018.06.03
HRCT has been used not only to diagnose pulmonary nodules such as malignant tumors but also to distinguish pre-invasive lesions (PILs) (AAH, AIS, or MIA) and IACs. In our hospital, the classification of GGNs by preoperative HRCT has become one basis for guiding resections of lung cancer.

The purpose of this study was to develop and validate a 3D CNN using quantitative imaging biomarkers obtained from HRCT scans to classify GGNs as PILs (AAH, AIS, or MIA) or IACs.

Methods

Patient selection

Ethical approval was obtained for this retrospective analysis, and informed consent requirement was waived. Our study recruited consecutive lung adenocarcinoma patients with pulmonary lesions (F/M, 1075/470; mean age 55.8±10.6 years; age range, 32–84 years) between January 2010 and March 2017. Within our institution’s PACS database, we selected patients with the words “GGN”, “GGO”, “ground glass opacity”, “ground glass nodule”, “non-solid nodule”, “sub-solid nodule” or “mixed nodule”. All patients underwent routine chest CT scans before surgery on their lung lesions. The inclusion criteria consisted of: (I) histopathologically confirmed PILs (AAH, AIS, MIA) and IACs; (II) the presence of solitary pulmonary nodule (with lesion measuring >5 and <30 mm in the longest diameter). Patients who met one of the following criteria were excluded: (I) previous systemic treatment (chemotherapy or radiotherapy), which could produce changes in texture features; (II) incomplete CT datasets at our institution; and (III) two or more lesions that had been resected.

CT acquisition and reconstruction

CT examinations including the chest and the upper abdomen were performed as part of the routine CT imaging protocol for the selected patients at our institution by using the Somatom Definition AS (Siemens Healthcare, Germany), Sensation 64 (Siemens Healthcare, Germany) and Brilliance (Philips Healthcare, the Netherlands), with a tube voltage of 120 kV and a current of 200 mA. The target lesion (the largest one if there were multiple nodules) was selected for reconstruction. All imaging data were reconstructed using a standard reconstruction algorithm, and the reconstruction parameters were as follows: slice thickness, 0.625 mm; increment, 1 mm; pitch, 1.078; field...
of view, 15 cm; and matrix, 512×512.

**Preprocessing**

First, all raw data were converted to Hounsfield units (HUs), a standard measure of radio-density in CT images. The preprocessing procedure consisted of two steps: lung mask extraction and the 3D reconstruction. Lung mask extraction removes tissues that are not in the lung but are similar to nodules in their spherical shapes, which reduces false positive signals in the detected lesions. The 3D reconstruction enables the developed algorithm to preprocess the data collected from different scanners or imaging protocols. For lung mask extraction, morphological operations (erosion and dilation) were utilized to calculate all connected components, and strategies (e.g., convex hull and dilation) were employed to generate the left and right lungs (17). Then, 3D reconstruction was performed by resizing the data to 1 mm × 1 mm × 1 mm resolution, and the luminance was clipped between −1,200 and 600, scaled to 0–255 and convert to unsigned 8 bits integer.

**Automatic lung cancer diagnosis**

Automatic diagnosis of lung cancer involves the detection of all possible nodules and classification of GGNs as PILs or invasive lung adenocarcinoma. A 3D CNN framework was designed for nodule detection and invasive lung adenocarcinoma probability evaluation. The flow of the automatic diagnosis framework is shown in Figure 1.

For better nodule detection performance, multi-resolution data was used to train the nodule detection network. For high resolution CT (HRCT) data (thickness = 0.625 mm), raw resolution helps to retain initial information. For thicker data (e.g., thickness ≥1.5 mm), resizing to unified resolution (e.g., 1 mm × 1 mm × 1 mm) improves nodule detection, particularly on CT data with collected with small thickness. Thus, the data with raw resolution and 1 mm × 1 mm × 1 mm resized resolution were used during the training process, and raw resolution was used to predict nodules on HRCT data. One problem that should be considered during the training process is that high-resolution CT data cause GPU memory exploration. Thus, small 3D patches were cropped from the preprocessed data. The size of each cropped patch is 128×128×128×1 (height × length × width × channel) in space. If the cropped patch goes beyond the lung scan, the excess of cropped patch over the scanned data was padded with the value of 170, which corresponding to the intensity of normal tissue in CT. The cropped patch of 3D CT data will be entered into the network during network training. The amount of the training set, the validation set, and the test set were 1,075, 270, and 200, respectively.

An encoder-decoder structure-based framework was designed to detect nodules in the input data. The encoder-decoder structure (18) enables the network to conveniently combine multi-scale resolution. A region proposal network proposed in Faster-R-CNN (19) was employed to first roughly locate the nodules and then to further regress the position and the diameter with the trained model. When training detection network, the learning rate was initialized...
as 0.01, and the learning rate decreased a factor of 10 every 50 epochs. Three different scales of square anchor boxes were used and the corresponding length with respect to the anchor boxes were 2, 10, and 30 mm, respectively. In the process of network design, key techniques, such as 3D convolution and hard-example mining, were used to promote the recall and accuracy of the nodule detection.

Based on the nodule detection results, the probability of one subject suffering invasive lung adenocarcinoma was evaluated with a shallow neural network. The classification network consisted of a 64-neuron, fully connected layer and a one-neuron output layer. The classification network was added as a branch in the decoder structure in the detection network. Accurate nodule detection enables convincing cancer classification; thus, detection and classification networks were alternately trained after the first 100 iterations. In this way, the information obtained from detection presented compelling evidence in making the final classification verdict.

Performance of radiologists with time constraints

An experienced radiologist (SPW, 14 years of medical imaging experience and 10 years of specialized chest imaging) labeled lung nodules on the test dataset and provided a list of nodule locations. Three radiologists volunteered to participate in the study. They understood and agreed with the basic principles and objectives of this study. As determined by the research ethics committee, the radiologists participating in the review panel were not required to provide written informed consent. Radiologists 1, 2, and 3 had 3, 7, and 15 years of experience, respectively. The average age of the three radiologists was 47.7 years (range, 32–45 years) and the average number of years of practice was 7.3 years (range, 3–15 years). Radiologist 3 had experience in chest imaging as a special interest area.

Three radiologists with time constraints (radiologists WTC) independently evaluated the corresponding CT slices of these nodules using our hospital’s daily diagnostic display (E-3620, MP Single-head Nio Grayscale Display, BRACO, Italy). The three radiologists were asked to express their confidence in the interpretation of each nodule according to five levels: definitely PIL, probably PIL, equivocal, probably IAC, and definitely IAC. To obtain an estimate of the sensitivity and specificity from each radiologist, the five confidence levels were split in two by considering the definitely PIL and probably PIL confidence levels as negative findings and all the other levels as positive findings. To obtain the ROC curve, the threshold was changed to cover the entire range of possible ratings by the radiologist, and the sensitivity was plotted as a function of false positive score (1-specificity). A performance comparison was conducted between the 3D CNN model and the three radiologists in the classification between PILs and IACs in CT using the same test data.

Statistical analysis

The results of our automatic lung cancer diagnosis algorithm were evaluated by comparing them to those of three radiologists with different years of diagnostic experience. The indicators of sensitivity, specificity, and accuracy were used to quantitatively assess the performance of the automatic algorithm. In addition, for clear comparisons with the results obtained by the radiologists, ROC curves and the AUC of the ROC were plotted (20). A higher AUC indicated a better invasive lung adenocarcinoma diagnosis result.

Results

Patient characteristics

Upon histological examination, 864 (55.9%) cases were PILs, and 681 (44.1%) cases were IACs. The mean age in the IAC group was significantly greater than that in the PIL group (P<0.01). The end-to-end network predicted the nodule position and diameter, together with the probability, with respect to the IACs. The statistical results listed in Table 1 indicate that 87.4% of the nodules classified as IACs had a diameter larger than 10 mm. In contrast, 87.7% of the nodules diagnosed as PILs had a diameter smaller than 10 mm. It is rather remarkable that the female patients made up a higher proportion of the subjects suffering from both PILs and IACs (P<0.01).

CNN performance

We used the AUC values of the overall classification accuracy, sensitivity, and specificity and the area under the ROC curve to evaluate the performance and discrimination of the deep learning model (CNN). Their quantitative measurements were calculated using accuracy, sensitivity, and specificity.

The ROC analysis results from the CNN model and from the three radiologists are shown in Figure 2. The
The CNN model had the largest AUC value of 89.2%, which was significantly different from the AUC of the three radiologists (radiologist 1: 80.5%; radiologist 2: 83.9%; and radiologist 3: 86.7%). Discriminant analysis showed that the CNN model was the best for the differential diagnosis of PILs and IACs. The total accuracy rates of radiologist 1, radiologist 2, radiologist 3, and the CNN model were 80.2%, 80.7%, 81.7%, and 84.0%, respectively. The sensitivity and accuracy of the CNN model were significantly higher than those of the three radiologists (P<0.05).

Discussion
This study demonstrated the deep learning algorithm was able to achieve a better AUC than a panel of three

Table 1 Clinicopathologic characteristics of the 1,545 patients in the study

<table>
<thead>
<tr>
<th>Variable</th>
<th>Final pathology</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Total (n=1,545)</td>
<td>Non-invasive (n=864)</td>
</tr>
<tr>
<td>Age, years (mean ± SD)</td>
<td>55.8±10.6</td>
<td>53.1±10.4</td>
</tr>
<tr>
<td>Sex, n (%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>1075 (69.6)</td>
<td>647 (74.9)</td>
</tr>
<tr>
<td>Male</td>
<td>470 (30.4)</td>
<td>217 (25.1)</td>
</tr>
<tr>
<td>Tumor diameter, cm, n (%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>≤1.0</td>
<td>912 (59.0)</td>
<td>758 (87.7)</td>
</tr>
<tr>
<td>1.1–2.0</td>
<td>551 (35.7)</td>
<td>89 (10.3)</td>
</tr>
<tr>
<td>&gt;2.0</td>
<td>82 (5.3)</td>
<td>17 (2.0)</td>
</tr>
<tr>
<td>Tumor location, n (%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RUL</td>
<td>542 (35.1)</td>
<td>336 (38.9)</td>
</tr>
<tr>
<td>RML</td>
<td>83 (5.4)</td>
<td>40 (4.6)</td>
</tr>
<tr>
<td>RLL</td>
<td>288 (18.6)</td>
<td>158 (18.3)</td>
</tr>
<tr>
<td>LUL</td>
<td>462 (29.9)</td>
<td>247 (28.6)</td>
</tr>
<tr>
<td>LLL</td>
<td>170 (11.0)</td>
<td>83 (9.6)</td>
</tr>
</tbody>
</table>

RUL, right upper lobe; RML, right middle lobe; RLL, right lower lobe; LUL, left upper lobe; LLL, left lower lobe; SD, standard deviation.
pathologists WTC were able to when participating in a simulation exercise involving the identification of PILs and IACs in stage IA lung adenocarcinoma. To our knowledge, this is the largest study based on HRCT and a deep learning algorithm to predict the invasive status of early lung adenocarcinoma so far. The sensitivity and accuracy of our GGNs classification model were 88.5% and 84.0%, respectively, which would be helpful in identifying IACs before operation. The results of our automatic classification method are competitive compared with other traditional methods (15), such as logistic regression (accuracy =81.5%), random forest (accuracy =83.0%), and adaboosting (82.1%).

Automatic interpretation of radiological image features of pulmonary nodules has potential benefits, such as improved efficiency, reproducibility, and improved prognosis, by providing early detection and treatment.

PILs included AAHs, AISs, and MIAs, and the disease-free survival rate after sublobectomy was 100% (4,12). However, lobectomy plus systemic lymph node dissection or sampling is the standard surgical approach for invasive adenocarcinoma (12). Given that intraoperative frozen sections are difficult to use to accurately determine the invasiveness of adenocarcinoma; preoperative identification of nodular invasiveness is helpful for preoperative planning (21). The CNN model predicted the invasiveness of GGNs with higher accuracy, sensitivity, and AUC values than those of the three radiologists.

A number of studies have reported the use of HRCT (22-26) or radiomics (27-29) in predicting early lung adenocarcinoma invasiveness. Methods such as HRCT and radiomics are limited by the significant pretreatment required to extract imaging biomarkers, reducing the reproducibility. Deep learning simplifies multistage pipelines by learning predictive features directly from the image, allowing for greater reproducibility. In this study, our accuracy (84.0%) was similar to or higher than that reported above, demonstrating that accurate predictions can be achieved in PIL and IAC patients without predesigned features. The sensitivity of our method is higher than those of the three radiologists. In fact, during the clinical diagnosis, the doctors tend to put more attention on patients suffered invasive lung cancer, which influences the value of sensitivity. Thus, our classification method is working on producing a high accuracy and high sensitivity. The results of the radiologists may be influenced by their experience. Some radiologists may prone to classify lung cancer as pre-invasive type, which produce a low sensitivity and a high specificity.

The algorithm was superior to the three time-limited radiologists in distinguishing the invasive status of lung adenocarcinomas. However, given sufficient time, radiologists may outperform the CNN algorithm in distinguishing PILs from IACs. In this study, the three radiologists were given only 3 hours to evaluate the CT images of all 230 nodules, each less than 1 minute. It is a remarkable fact that the locations of all nodules were provided for the three radiologists before reading. Thus, the radiologists only need to draw their attention on distinguishing whether a nodule is pre-invasive or invasive. There were two problems with this approach: one was that this was an impractical assessment of short-term CT images, and the other was that, in routine clinical practice, radiologists are unlikely to review 230 consecutive pulmonary nodules. In most hospitals, radiologists require additional image post-processing and windowing techniques, which are factors that affect diagnostic performance. In addition, the simulation exercise invited the radiologists to review 230 nodular CT sections within approximately 3 hours to distinguish between PILs and IACs. Although feasible in this simulation environment, this does not represent the progression of work in other settings. Reducing the time limit for task completion can improve the accuracy of GGN diagnostic review. Although there is no perfect control, it is clear that there were limitations on this study.

In fact, the 3D CNN-based automatic method made an advance in lung cancer diagnosis in comparison with human diagnosis. First, on the basis of the 3D convolution technique, the CNN algorithm makes full use of information from sagittal, coronal, and axial views. In this way, the automatic diagnosis algorithm obtains more comprehensive and accurate information with respect to nodule size, shape, and location, which improves the reliability of the diagnostic results. Second, as in human beings, CNN works as memory network by learning critical features in cancer classification. In this process, CNNs are more flexible in adjusting learning strategies and are able to determine the critical clinical features even for confusing cases. In contrast, radiologists may be limited by experiences, and the results are likely to be tendentious at times. For example, the results in Table 2 indicate some radiologist tendencies to classify IACs as PILs.

A limitation of our developed automatic lung cancer diagnosis method is the effect of an insufficient data size on
final performance. To compensate for the limitation, various data enhancement strategies, such as rotation, enhancement, and scaling, were employed during the training process. In the future, data collection should be continued for further enhancement of classification sensitivity and specificity with respect to IACs and PILs, respectively.

The focus on this work in the future will be to use a larger network in the hope of achieving more accurate results. Deeper networks can potentially improve the model's results and generalization capabilities; however, more multicenter data will be needed to minimize the effects of overfitting. A larger patient population will further improve the performance of the algorithm, especially considering the heterogeneity of the image acquisition parameters. The rates of IACs in our study varied among different patient groups. Ideally, all patient CT scans will be acquired using consistent acquisition parameters (kV, mAs, resolution, slice thickness, and contrast agent), and the rates of IACs will be the same. However, this will be a challenge in practice because the CT scanner model and acquisition parameters, as well as the captured patient's demographic data, vary widely among different institutions. Our research validated the algorithm by using data from a single institution, and we hope to use data from multiple institutions in the future to enhance the model's generalization capability.

**Conclusions**

In this paper, we have shown that the performance of a deep CNN trained on large datasets to classify IACs from PILs is superior to that of experienced radiologists. Our model may be used as a non-invasive tool to supplement invasive tissue sampling, to guide patient surgery protocols and to manage the disease early and during follow-up. We believe that deep CNNs or similar deep learning methods show great advantages, and these algorithms can ultimately help doctors identify PILs and IACs.

**Acknowledgements**

**Funding:** This research is supported by, Shanghai Science and Technology Funds (No. 20130424), NSFC (No. 81371550). We thank engineers from Tencent Youtu Lab for debugging automatic diagnosis algorithm.

**Footnote**

**Conflicts of Interest:** The authors have no conflicts of interest to declare.

**Ethical Statement:** Ethical approval was obtained for this retrospective analysis, and informed consent requirement was waived.

**References**

5. Moon Y, Lee KY, Park JK. The prognosis of invasive

---

**Table 2** Diagnostic performances of the CNN model and the three radiologists

<table>
<thead>
<tr>
<th>Model</th>
<th>AUC (%)</th>
<th>Accuracy (%)</th>
<th>Sensitivity (%)</th>
<th>Specificity (%)</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Radiologist 1</td>
<td>80.5</td>
<td>80.2</td>
<td>71.4</td>
<td>86.7</td>
<td>&lt;0.01</td>
</tr>
<tr>
<td>Radiologist 2</td>
<td>83.9</td>
<td>80.7</td>
<td>79.8</td>
<td>81.4</td>
<td>&lt;0.01</td>
</tr>
<tr>
<td>Radiologist 3</td>
<td>86.7</td>
<td>81.7</td>
<td>64.2</td>
<td>94.7</td>
<td>&lt;0.01</td>
</tr>
<tr>
<td>Our method</td>
<td>89.2</td>
<td>84.0</td>
<td>88.5</td>
<td>80.1</td>
<td></td>
</tr>
</tbody>
</table>

Statistically significant difference from our method. CNN, convolutional neural network; AUC, area under the curve.